THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

# Comparison of models for estimation of long-term exposure to air pollution in cohort studies

**Citation for published version:**
Beverland, IJ, Robertson, C, Yap, C, Heal, MR, Cohen, GR, Henderson, DEJ, Hart, CL & Agius, RM 2012, 'Comparison of models for estimation of long-term exposure to air pollution in cohort studies' Atmospheric Environment, vol 62, pp. 530-539. DOI: 10.1016/j.atmosenv.2012.08.001

**Digital Object Identifier (DOI):**
10.1016/j.atmosenv.2012.08.001

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Peer reviewed version

**Published In:**
Atmospheric Environment

**Publisher Rights Statement:**
Author's Post-print: author can archive post-print (ie final draft post-refereeing)

**General rights**
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 28. Apr. 2017

**Post-print of peer-reviewed article published by Elsevier.**
**Published article available at: http://dx.doi.org/10.1016/j.atmosenv.2012.08.001**

**Cite as:**
**Beverland, I.J., Robertson, C., Yap, C., Heal, M.R., Cohen, G.R., Henderson, D.E.J., Hart, C.L. and Agius, R.M. (2012) Comparison of models for estimation of long-term exposure to air pollution in cohort studies, *Atmospheric Environment* 62, 530-539.**

# Comparison of Models for Estimation of Long-Term Exposure to Air Pollution in Cohort Studies

Beverland, I.J.*
*Department of Civil Engineering, University of Strathclyde, Glasgow, UK*

Robertson, C.
*Mathematics & Statistics, University of Strathclyde;*
*Health Protection Scotland, Glasgow, UK; and*
*International Prevention Research Institute, Lyon, France*

Yap, C.
*Department of Civil Engineering, University of Strathclyde, Glasgow, UK*
*(Current address: MRC Midland Hub for Trials Methodology Research, University of Birmingham, UK)*

Heal, M.R.
*School of Chemistry, University of Edinburgh, Edinburgh, UK*

Cohen, G.R.
*Edinburgh, UK*

Henderson, D.E.J.
*Department of Civil Engineering, University of Strathclyde, Glasgow, UK*
(Current address: SKM Enviros, Shrewsbury, UK)

Hart, C.L.
*Institute of Health and Wellbeing, Public Health, University of Glasgow, Glasgow, UK*

Agius, R.M.
*Centre for Occupational and Environmental Health, The University of Manchester, UK*

***Correspondence to:**
Dr Iain Beverland,
Senior Lecturer,
University of Strathclyde, Department of Civil Engineering,
John Anderson Building, 107 Rottenrow,
Glasgow G4 0NG
Tel: 0141-548-3202
Fax: 0141-553-2066
Email: Iain.Beverland@strath.ac.uk

Highlights:

- Three approaches for modelling long-term exposure to black smoke were evaluated
- Missing data require development of imputation procedures
- Cross validation & GIS-based visualisation were used for model evaluation
- Marked differences in performance were noted in evaluation of different models
- Improved black smoke exposure estimates were observed using a multi-level model

**Abstract:**

This study compared three spatio-temporal models for estimation of exposure to air pollution throughout the central part of Scotland during 1970-79 for approximately 21,600 individuals in 2 closely-related prospective cohort studies. Although 181 black smoke (BS) monitoring sites operated in this region at some point during 1970-79, a substantial amount of BS exposure data was missing at many sites. The three exposure estimation methods were: (i) area-based regression models to impute missing data followed by assignment of exposure by inverse distance weighting of observed BS at nearby monitoring sites (IDWBS); (ii) area-based regression models to impute missing data followed by a spatial regression additive model using four local air quality predictors (LAQP): altitude; distance to the nearest major road; household density within a 250 m buffer zone; and distance to the edge of urban boundary (AMBS); (iii) a multilevel spatio-temporal model using LAQP (MultiBS). The three methods were evaluated using maps of predicted BS, and cross validation using monitored and imputed BS at sites with ≥ 80% data. The use of LAQP in the AMBS and MultiBS exposure models provided spatial patterns in BS consistent with known sources of BS associated with major roads and the centre of urban areas. Cross-validation analyses demonstrated that the MultiBS model provided more precise predictions ($R^2$ = 60%) of decadal geometric mean BS concentrations at monitoring sites compared with the IDWBS and AMBS models ($R^2$ of 19% and 20%, respectively).

## 1. Introduction

Epidemiological research in cohort studies has found consistent associations between long-term exposure to air pollution and adverse health effects (Brunekreef and Holgate, 2002; Pope and Dockery, 2006). A crucial challenge for such research is to reliably assign estimates of air pollution exposure to the individuals being studied. Earlier analyses in the United States on the 'Six Cities' study (Dockery et al., 1993) and the American Cancer Society (ACS) study (Pope et al., 2002; 1995) relied on inter-urban variations in air pollution with discrete urban areas represented by single monitoring sites. Subsequent attention has focused on exposure assignment determined from intra-urban variations in pollutant concentrations (Krewski et al., 2009). For example, substantially higher pollution-mortality associations have been found in the Los Angeles subset of the ACS study using estimates of marked intra-urban variations in exposure interpolated from multiple monitoring sites (Jerrett et al., 2005). Other studies have used land-use regression and/or dispersion modelling approaches to estimate small area variations in pollutant exposure taking into account proximity to road traffic (Beelen et al., 2010; Gulliver and Briggs, 2011; Gulliver et al., 2011; Hoek et al., 2008) and geo-statistical exposure estimation using spatio-temporal models (Dadvand et al., 2011; Fanshawe et al., 2008; Gryparis et al., 2007; Yanosky et al., 2008).

The aim of the work described here was to develop and compare methodologies to estimate long-term exposures to air pollution for individuals in two closely related cohort studies in Scotland. The scientific challenges in this study and how they were tackled include: incomplete exposure data requiring evaluation and comparison of innovative exposure modelling approaches (as outlined here); application of novel exposure estimation methodologies and examination of impacts of historical pollution over a long follow up period (Yap et al., In press); and novel comparison of pollution-outcome associations at different exposure timescales (Beverland et al., 2012). A specific objective was to estimate exposure to black smoke (BS) over a 10 year period in the 1970s when the cohorts were recruited. BS is a metric of the optical darkness of airborne particulate matter collected on filter media (Heal and Quincey, 2012). Although quantified in units of $\mu g\ m^{-3}$ BS concentrations do not equate directly to the mass of airborne particulate matter. However, consistent standard calibrations (e.g. DETR (1999)) have been used for many decades to convert reflectance to nominal concentration such that BS data are important measures of historic levels of air pollution used widely in epidemiological studies. The DETR (1999) calibration procedures were used in the computation of UK government archived BS data used in the present manuscript. The use of the BS metric as a measure of particulate matter air pollution is well-established in the epidemiological research community and has been shown to be a good marker for traffic and other primary

combustion-related urban air pollution, often at least as predictive of negative health outcomes as $PM_{10}$ or $PM_{2.5}$ (Janssen et al., 2011).

Although, there was a relatively large number of BS air pollution monitoring sites in operation in the cohort areas for at least part of the 1970s, we faced problems of missing values in the recorded data and uneven spatial distribution of monitoring sites and cohort individuals. To overcome consequent challenges in attributing exposure to individuals on the basis of their place of residence we developed, in addition to relatively simple imputation methods, area and spatial regression models, and more complex multilevel models, using geographical local air quality predictors (LAQP) derived using a Geographical Information System (GIS).

## 2. Methods

### 2.1 Geographical coverage of exposure models:

The models described here were developed to estimate BS concentrations across central Scotland, for application to air pollution epidemiological analyses of two of the Midspan Cohorts (Hart et al., 2005). The geographical coverage of the locations of cohort individuals is shown in supplementary information Figure S1. The Collaborative Cohort, recruited in 1970-73, consisted of approximately 7,000 employed individuals, aged 35-64, geographically dispersed throughout the cities of Glasgow and Edinburgh and towns and villages in the central part of Scotland (Supplementary Information Fig S1). The Renfrew/Paisley Cohort consisted of approximately 15,400 individuals, aged 45-64 during 1972-76, from the towns of Renfrew and Paisley in West Central Scotland. As an indication of geographical scale the contiguous conurbation of Glasgow, Paisley and Renfrew can be encompassed within a radius of 12 km; with Renfrew and Paisley encompassed within circles of radii 1.5 and 3.5 km respectively within this 12 km radius. Results from the cohort analyses are given elsewhere (Beverland et al., 2012; Yap et al., In press).

### 2.2 Air pollution observations

BS exposure was estimated for each individual in both cohorts for 1970-79 to include the baseline data collection periods and short periods afterwards. BS data from 181 monitoring sites operating in the central part of Scotland at some point during this decade (Figure S1) were obtained from the UK National Air Quality Information Archive (DEFRA, 2005). The location and periods of operation of the sites were at the discretion of local and national authorities responsible for air pollution. These BS measurements were made at a time when there was a move away from using coal as the main source of domestic heating fuel.

Examples of the nature of the missing data are illustrated in Section S2 supplementary information. Only 24 out of 181 sites had data available for > 90% of days in 1970-79; only 67 sites had > 60% of data; and 19 sites had < 10% of the potentially available data. As the amount of missing data was substantial and the data were not missing at random, we developed models to impute the missing values and to assign geometric mean exposure estimates to individuals' residential addresses. In the period from 1980 onwards there was a substantial reduction in the number of sites which recorded BS and consequently we could not pursue our analyses to estimate exposures during this later period.

## 2.3 Exposure modelling

### 2.3.1 Overview

Three types of exposure models were developed. The first approach grouped monitoring sites within geographical regions and used a log-linear regression model to impute missing daily BS observations. BS concentrations at individuals' residential addresses were estimated from inverse-squared distance weighted averages of BS at nearby monitoring sites. This model is abbreviated as 'IDWBS.' The second method used daily observations and imputations from the first (area-based) method in a semi-parametric model that replaced the distance weighted assignment with an additive model with LAQP. This additive model is abbreviated as 'AMBS.' The third method, involving a multilevel spatio-temporal model of monthly BS, referred to as 'MultiBS', had the capability to estimate coefficients for the LAQP in the presence of missing data, thus allowing predictions of the missing observations to be made from the fitted model.

### 2.3.2 Area-based log-linear regression and inverse distance weighted exposure model (IDWBS)

Using local knowledge of the geography and meteorological conditions in Scotland monitoring sites were grouped in 15 geographic regions (Figure S1), with a criterion that each region had ≥ 1 site with ≥ 60% available BS data. The following model was used to impute missing data and compute geometric mean daily BS for 1970-79 at sites within each region:

$$y_{ij} = ln\left(BS_{ij} + 0.5\right) = s_i + \beta_1 t_{ij} + \beta_2 t_{ij} I\left(t_{ij} \rangle\, t^*\right) + day\left(t_{ij}\right) + month\left(t_{ij}\right) + \varepsilon_{ij} \qquad (1)$$

where $i$ indexes site, $j$ indexes the sequential observations for a site, $BS_{ij}$ is black smoke on observation $j$ at site $i$, $t_{ij}$ is time (days from 1/1/1970), $I\left(t_{ij} \rangle\, t^*\right)$ is an indicator variable to indicate if time is after decadal midpoint of 1/1/1975, $day(t_{ij})$ and $month(t_{ij})$ are indicators for weekday and calendar month, $s_i$ is a site specific intercept, and $\varepsilon_{ij}$ is an error term. Common piecewise linear trend, daily and seasonal effects were assumed for all the monitoring sites within each region, but these were allowed to vary between regions.

Although this model was too simple to describe the true complexity in daily BS observations, it provided a description of general trends within each region. Equation 1 produced broadly equivalent estimates to more complex imputation models (Section S3 supplementary information provides further detail of model evaluations).

BS exposure, $BS_x$, at each individual's address at location $x$ was assigned using inverse distance weighting of the decadal geometric-mean $BS_i$ for site(s) $i$ calculated using the imputation of Equation 1:

$$BS_x = \frac{\sum_{i=1}^{n_x} \frac{BS_i}{\left(d_{ix}^*\right)^2}}{\sum_{i=1}^{n_x} \frac{1}{\left(d_{ix}^*\right)^2}} \qquad (2)$$

where $n_x$ is number of sites within 1 km of $x$ and $d_{ix}^*$ is distance between site $i$ and $x$. If there was only one site within 1 km then $BS_x = BS_i$. If there were no sites within 1 km, exposure was assigned the weighted average of the 2 nearest monitoring sites regardless of distance. The entire Renfrew-Paisley cohort lived within 5 km of a monitoring site as did 95% of the Collaborative Cohort. In the IDWBS method, cohort individuals could not have an exposure greater than the maximum, or less than the minimum, of the 2 nearest sites.

2.3.3 Local Air Quality Predictors (LAQP) and the additive exposure model (AMBS)

The IDWBS method of assigning exposure to an individual does not allow for environmental determinants of local BS concentrations. It is anticipated that in the 1970s the predominant source of BS in the study area was household coal fires, though there would have been some contribution from car exhaust fumes, industrial sites and power stations. There is likely to have been geographic variation in pollutant dispersal associated with altitude, which affects wind speed, and advection of clean air from areas outside the urban boundaries. We therefore modelled the relationship between average *ln* BS (calculated after imputation using the area-based log-linear regression model, Equation 1) and LAQP associated with the above emission, dispersion and advection processes, using linear regression (parametric) models and additive (semi-parametric or nonparametric) models.

The following LAQP variables were calculated at postcode centroids of individual addresses using GIS procedures (Table 1): easting (*E*), northing (*N*), altitude (*A*) (Ordnance-Survey, 1993); distance to nearest major road (*MR*) (motorways and 'A' roads in 2001 (NAEI, 2005)); household density in a 250 m buffer (*HD*) (calculated from 1981 census data (SURPOP, 2006)); distance to nearest edge of the urban boundary (*UB*) (Ordnance-Survey, 2003); and an indicator for whether the postcode centroid was inside or outside of small (< 17.7 $km^2$, cut-point defined by median area of urban areas containing monitoring sites) or large urban areas (*UA.Ind*). Maps illustrating LAQP are provided in Section S4 supplementary information. Distances to the nearest major road and traffic intensity were highly correlated. Both of the variables were only available in 2001, but not in earlier years.

The distance to the nearest major road was preferred as there were less likely to have been major changes to the locations of the major roads, compared to substantial traffic flow changes between 1970-79 and 2001.

Five spatial regression models were examined in sensitivity analyses (details in Section S5 supplementary information). These ranged from a fully flexible additive model to a linear regression model with no spatial smoothing. The most parsimonious configuration was a semi-parametric model with bivariate smooth trend, *s(E,N)* and parametric terms for LAQP:

$$\ln(BS+0.5) = s(E, N) + \beta_1 A + \beta_2\sqrt{HD} + \beta_3\sqrt{MR} + \beta_4 UB + \beta_5 UA.Ind \quad (3)$$

where $\beta_1 \ldots \beta_5$ are fixed effects parameters for LAQP. This model gave a high $R^2$ of 75%, with the lowest Bayesian Information Criterion (BIC), and second lowest Akaike Information Criterion (AIC) (Table S3 supplementary information). Normality and linearity assumptions were valid for this model. We investigated the presence of spatial correlation in mean *ln* BS concentrations. An exponential variogram was fitted using non-linear weighted least squares using *variogram.fit()* in the *SpatialStats* module of Splus 7.0. The estimated range of dependence was 2 km and the sill was 0.069 implying that the spatial correlation was negligible for this model.

Mean *ln* BS increased with increasing household density and distance from the urban edge (i.e. closer towards the centre of the urban area); and decreased with increasing altitude and increasing distance from a major road. This was consistent with *a priori* expectations of increased BS in areas with greater household density (more coal fires and more local traffic emissions) and proximity to the centre of urban areas (more traffic and less advection of clean air masses) and major roads (traffic emissions); and decreased BS at higher altitudes (increased wind speed and dispersion).

2.3.4 Multilevel model (MultiBS)

The AMBS analysis using LAQP described above was a two-stage process of imputation followed by estimation. We also investigated the use of a multilevel model to simultaneously model the temporal change in BS at all sites and the spatial variation in LAQP across sites. This allowed subsequent estimation of BS exposures at individual addresses using LAQP. Multilevel modelling provides an estimate of the between-site variability. If observations within a single site over time are more similar compared to observations from other sites, then not taking into account this hierarchical structure present in the data would result in incorrect standard errors, and loss of information about between-site variability.

We used a semi-parametric mixed-effects model to fit site-specific curves with random coefficients. The model described *ln* BS at each site as the sum of a time-dependent population mean, and a site-specific deviation from the population mean. Both the population mean and site-specific deviations were modelled in a semi-parametric way, using penalised linear splines with random coefficients (Durban et al., 2005). Fixed effects for day of week, month, and LAQP were included. Penalised linear splines were used to reduce the computational load associated with the large data-sets involved. As there were 181 sites with potential data over 10 years, daily data were too cumbersome to deal with and with very little loss of information we developed more efficient models using monthly averaged data. Data from a month was included provided there were at least 15 daily observations.

In the fixed-effects model (Equation 4), $y_{ij}$ denotes the monthly mean of $ln(BS_{ij}+0.5)$ of site $i$ at time $t_{ij}$ (number of months from January 1970), where $i = 1\ldots s$ indexes the sites, and $j = 1\ldots n_i$ indexes the time series of observed monthly data (not necessarily consecutive months) within site $i$, $n_i$ is the number of months of data available for site $i$:

$$y_{ij} = f\left(t_{ij}\right) + g_i\left(t_{ij}\right) + \alpha_{1c} Cos\left(\frac{t_{ij}}{12}\right) + \alpha_{1s} Sin\left(\frac{t_{ij}}{12}\right) + \alpha_2 A_i + \alpha_3\sqrt{HD_i} + \alpha_4\sqrt{MR_i} + \alpha_5 UB_i + \varepsilon_{ij} \qquad (4)$$

$f(t_{ij})$ is a parametric or nonparametric function representing the decreasing trend of BS over time, averaged over the population of all sites and $g_i(t_{ij})$ represents the deviation of the $i^{th}$ site from the population mean at time $t_{ij}$. Sine and cosine terms model seasonal effects. The fixed-effect parameters $\alpha_{1c}$, $\alpha_{1s}$ represent monthly seasonality; $\alpha_2 \ldots \alpha_5$ represent effects of LAQP. Preliminary analyses found that the urban area indicator was not an important predictor variable and so was not included. The 'within-site' error term, $\varepsilon_{i,j} \sim N\left(0,\sigma^2\right)$, is a random variable representing the deviation of predicted mean *ln* BS for month $j$ at site $i$ from observed *ln* BS. Easting and Northing were considered at a second stage.

The population trend, $f(t_{ij})$, was estimated using a penalised linear spline (Durban et al., 2005):

$$f\left(t_{ij}\right) = \beta_0 + \beta_i\, t_{ij} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ \qquad (5)$$

where $\kappa_1 \ldots \kappa_K$ was a set of distinct knots between $t_{ij}$ and $t_+ = max(0, t)$. The use of $K = 9$ knots allowed the slope for each year to differ. The change in slope at the knots was a random effect, $u_k \sim N\left(0,\sigma^2\right)$. A larger number of knots would have ensured greater flexibility, but would have increased computational load. Investigations showed that more flexibility was unnecessary.

Estimates of $\beta$ and $u_k$ were obtained by minimizing the penalized least squares (Durban et al., 2005).

The simplest way of representing site-specific differences is to include a random effect for each monitoring site:

$$g_i(t_{ij}) = V_i \qquad (6)$$

where $V_i \sim N(0,\sigma_V^2)$ denotes deviation of the $i^{th}$ site from the population trend. In this model deviation of each site from the overall trend is modelled as a random intercept, which assumes that trends for different sites are parallel. As this assumption may be in doubt, an extension of modelling site-specific differences as random intercepts allowed the underlying linear trend to vary over sites.

$$g_i(t_{ij}) = b_{i1} + b_{i2}\, t_{ij} \qquad (b_{i1},b_{i2}) \sim N(0,\Sigma) \qquad (7)$$

A further extension was to model $g_i(t_{ij})$ using penalised linear splines to allow linear variations in slope in each year to vary randomly across sites,

$$g_i(t_{ij}) = b_{i1} + b_{i2}\, t_{ij} + \sum_{k=1}^{K} v_{ik}(t_{ij} - \kappa_k)_+, \qquad (b_{i1},b_{i2}) \sim N(0,\Sigma) \quad v_{ik} \sim N(0,\Sigma) \qquad (8)$$

where $v_{ik}$ coefficients represent slope deviations.

In sensitivity analyses parameter estimates for different model configurations were obtained using Restricted Maximum Likelihood methods (Section S6 supplementary information). A model defined by Equations 4, 5 and 8 with an additional term to model the serial correlation between data for successive months (coded ML3-IT(NP-Serial) in Section S6) had the highest log likelihood and the lowest AIC and BIC (Table S4), and was therefore used in subsequent analyses. The parameter estimates for ML3-IT(NP-Serial) showed that BS at monitoring sites decreased with increasing altitude and with increasing distance from a major road, and that BS increased with increasing household density and with proximity to an urban centre (Table S5), consistent with expectations. Predictions from this model are presented for four example sites in Figure 1. Temporal trends differed over the sites, as did the average concentrations which were influenced by LAQP, but the seasonal (monthly) effects were constrained to be the same at all sites.

With the multilevel model all 181 monitoring sites had 120 predicted monthly means of *ln* BS i.e. missing BS data were effectively imputed. It would have been possible to retain the original data where it was available and only impute months which were missing, as was done for the log-linear model on daily data (Equation 1), but this was not done as the predicted values included the random effects.

Estimates of *ln* BS at cohort individual addresses were derived from those at monitoring sites by the following three-stage process. Firstly, monthly average *ln* BS values for each site was predicted by substituting the mean values for altitude, household density, major road distance and urban boundary distance over all 181 sites, and the 10-year averages of *ln* BS for each site calculated. The differences in these 10-year average *ln* BS between sites were thus due only to the estimated random intercepts and temporal trends for the sites. The single site-specific values at this first stage were effectively 'residual' effects adjusted for the LAQP. In the second stage estimated *ln* BS was extrapolated to cohort individuals by imputing values for these 'residual' effects at individual addresses by spatially smoothing the 'residual' effects at the 181 sites using bivariate splines, with 25 degrees of freedom. This gave a map of predicted mean *ln* BS at any point in central Scotland, assuming all points had the same altitude, household density, distance to the nearest major road and distance to the urban boundary. In the third stage LAQP were calculated for each cohort individual location. The fixed effects were then added for each individual, using the parameter estimates $\alpha_2 \ldots \alpha_5$ from the MultiBS model (Table S5). Thus BS was estimated for each cohort individual based upon the combination of the LAQP at their household address and the smoothed 'residual' effects of the sites around each address location.

2.4 Exposure model evaluation

All sites with ≥ 80% data (39 sites) were selected and missing data imputed with a site-specific time-series model with a flexible trend, month and day effects:

$$y_{ij} = f_i\left(t_{ij}\right) + \sum_{d=2}^{7} \sigma_{i1d}\gamma_d\left(t_{ij},d\right) + \sum_{m=2}^{12} \sigma_{i2m}\gamma_m\left(t_{ij},m\right) + \varepsilon_{ij} \qquad (9)$$

where $y_{ij} = \ln\left(BS_{ij} + 0.5\right)$ for site $i$ at time $t_{ij}$ (measured in days from 1/1/1970); with $i$ = 1,... $s$ indexing the sites and $j$ = 1,... $n_i$ indexing the $n_i$ observations on site $i$. This was similar to the first part of the multilevel model (Equation 4) but used daily data and a generalised additive model based upon a normal distribution with a smoothing spline, which had a target of 20 degrees of freedom as opposed to the piecewise linear spline with 9 knots. Furthermore, this imputation model included indicator terms for the day of the week and month of the year, where $\gamma_m\left(t_{ij},m\right) = 1$ if $t_{ij}$ is in month $m$ and 0 otherwise, and $\gamma_d\left(t_{ij},d\right)$ defined similarly for days. Missing values in the original data series were imputed using predictions from this model with addition of a simulated value from a normal distribution with mean zero and variance $\hat{\sigma}_i^2$ (which is the estimated variance of the residual term $\varepsilon_{ij}$).

Half of these 39 sites (i.e. 19 sites) were selected as 'test' data, leaving 162 sites as 'training' data (20 unselected sites with complete data plus 142 sites with < 80% available data). Using the training data we fitted the three exposure models and predicted average $ln(BS+0.5)$ over the ten year period for the 19 test sites. Ten separate random selections of the 19 test sites and 162 training sites were made and explained variance ($R^2$), root mean square error (RMSE), bias and fractional bias (FB) between the 190 predicted and 'test' exposures were calculated.

## 3. Results

### 3.1 Visualisation of estimated BS spatial variations:

Figures 2 and 3 show maps of predicted BS concentrations derived from the IDWBS, AMBS and MultiBS exposure models produced by kriging spatial smoothing within ARCGIS of the BS estimates from each model at the address postcodes of the 21,621 cohort individuals. These maps were consistent with those from an alternative application of a nonparametric bivariate smooth trend using penalised thin plate splines (Wood, 2003) fitted using the *mgcv* package in R (R-Development-Core-Team, 2006) (supplementary information section S7, Figure S4). The only purpose of the secondary smoothing in Figs 2, 3 and S4 was to provide the visual representations of the BS concentrations. The concentrations on these maps were not used in epidemiological analyses.

In predictions of BS over central Scotland (Collaborative Cohort) the IDWBS estimated concentrations were notably high in an implausibly extensive area extending tens of kilometres beyond the urban boundaries of former mining (therefore coal burning) towns and villages to the north east of Glasgow, with relatively limited extent of high concentrations within central Glasgow (Figure 2(C) and Edinburgh (Figure S4(C)). The highest AMBS concentrations were predicted in central Glasgow (Figure 2(B)), with high predicted concentrations also to the north east of Glasgow and in Edinburgh (Figure S4(B)). Predicted concentrations for all methods were lower in the rural areas especially those to the south and west of Glasgow, and on elevated ground between Glasgow and Edinburgh. These patterns were in accordance with the prevalent wind directions over central Scotland. The MultiBS model produced similar overall predictions to the AMBS model, although without the high concentrations to the north east of Glasgow (Figures 2(A) and S4(A)).

In the Renfrew-Paisley Cohort sub-area the AMBS and MultiBS models provided a much more consistent and plausible prediction of exposure at addresses of individuals than the IDWBS model (Figure 3). Overall patterns of modelled AMBS and MultiBS concentrations were similar but the concentrations were slightly higher for MultiBS. The estimated AMBS and MultiBS concentrations were highest in the centre of Paisley and decreased with distance from the town centre, with pollution contours tending to follow the major roads. The land rises to the south west of Paisley town centre and this increasing altitude contributed to the lower predicted BS concentrations in this predominantly residential area with few major roads (where local knowledge anticipated relatively low pollution concentrations). The IDWBS model failed to predict these anticipated lower concentrations in SW Paisley as all estimates were constrained to remain within the high concentrations measured at the nearest sites in the centre of Paisley.

3.2 Exposure model statistical evaluation

The MultiBS model had a substantially higher $R^2$, and lower RMSE and root mean square bias than the other two models, albeit with some evidence of overestimation of the test $ln(BS+0.5)$ data (Table 2).

4. Discussion

This manuscript describes the use of incomplete air quality data, collected for the administrative purposes of monitoring the effectiveness of smoke control measures and compliance with air quality standards, for application to epidemiological analyses of two cohorts. The main differences between the three exposure models developed here were the way in which spatial and temporal trends were modelled, the use of LAQP, and estimation of effects of LAQP before or after imputation. The MultiBS and AMBS models both made use of spatial smoothing; the main differences between them were more flexible linear trend prediction in the MultiBS model and imputation before estimation in the AMBS model. The difference between AMBS and IDWBS models was the use of LAQP and spatial smoothing in the former as opposed to distance to nearest monitoring sites in the latter.

As RMSE for AMBS and IDWBS models were similar, the addition of spatial smoothing and LAQP did not substantially improve predictions at monitoring sites. However, as the IDWBS method is based upon a pre-specified division of Scotland into 15 geographic regions (Figure S1) there is indirect partial inclusion of spatial predictor variables in this method (easting and northing especially and to a lesser extent altitude). The IDWBS approach is not likely to be as useful as the other two methods for estimating exposure at household locations. Many of the sites with long-term data were in similar urban areas where relative changes in LAQP from site to site were relatively small which is likely to have mitigated the effect of omission of these variables in model evaluation. Most monitoring sites were in the centre of towns and close to main roads. All households within 1 km of such sites would be assigned identical or similar BS concentration even though the house may be in a residential area distant from the main road. An equivalent limitation applies to spatial variations in household density. These limitations of the IDWBS method are manifested in some of the unexpected and implausible patterns of exposure noted in the maps of predicted exposures (Figures 2-3).

The AMBS and MultiBS models provided consistent and more plausible predictions of spatial patterns of exposure at addresses of individuals (Figures 2-3) as the predicted concentrations were not constrained to lie within the range of concentrations at the monitoring sites by making allowance for the important influence of LAQP. LAQP were significantly associated with the

concentrations of BS within the spatial regression models used. Using these predictors, which are known for each postcode location, is likely to improve the prediction of individual exposure at place of residence. Of course we have no way of directly verifying this but we have conducted sensitivity analyses to investigate the effects of our choices. The model evaluations suggested further benefits of MultiBS estimation of LAQP effects using only observed data and in better estimation of pollutant trend over time, which varies between sites. For example the data in Table 2 show that the MultiBS model evaluation data had substantially lower dispersion (lower root mean square difference and lower standard deviation of bias) than the AMBS (and IDWBS) model.

The models that we have developed have similarities to those of Maynard et al. (2007) who used a smoothed fixed-effect model for predicting daily black carbon in Boston, and Yanosky et al. (2008) who used a two-stage spline approach to model $PM_{10}$ in the United States. Shaddick and Wakefield (2002) used a Bayesian model of daily multiple-pollutant data at 8 sites to predict concentrations at individual households within the same model, whereas we adopted a two-stage process. A two-stage approach was also taken by Dadvand et al. (2011) and Fanshawe et al. (2008) for modelling black smoke in North East England. Sahu et al. (2006) used Monte Carlo Markov Chain (MCMC) techniques to predict individual exposures allowing for space and time dependence. Our approaches have similarities to the MCMC approach, though have less spatial dependency; our investigations of the generalised additive model revealed little spatial correlation and the multilevel model allows for temporal correlation.

Other methods of estimating BS over a geographic area include the use of dispersion and/or land-use regression (LUR) models (Beelen et al., 2010; Gulliver and Briggs, 2011; Gulliver et al., 2011; Hoek et al., 2008). Dispersion models take into account the magnitude of specific sources of the pollutants. This approach was not utilised as geographical variations in emissions from the main sources of BS in the 1970s (household fires and traffic) are poorly characterised. Proxies for these sources were used in our exposure models. Our use of LAQP is similar to that employed in LUR models and share common limitations associated with geographical accuracy (e.g. through uncertainties in exact locations of road links and of households in relation to postcode centroids) and simplifications within the exposure models (e.g. not allowing for irregularly shaped urban areas when computing distance to urban edge variable).

We identified a limited number of studies with similar BS exposure models. The most comparable study involves LUR model estimates of BS at 1 km × 1 km spatial resolution for all of the UK, including model evaluation using an independent set of 20% of ‘hold out’ monitoring sites (Gulliver

et al., 2011). Gulliver et al. observed deterioration in LUR model performance between 1971 and 1981 and attributed this to reduced variation in monitored concentrations. Correspondingly we compared our model evaluation statistics for 1970-79 by computing averages of Gulliver et al.'s broadly equivalent statistics for 1971 and 1981 and noted that our regional-scale multi-level model compared favourably with this national-scale LUR (e.g. $R^2$ values for obs-mod *ln* BS of 60% and 59%, respectively) but our AMBS and IDWBS models had notably poorer performance statistics ($R^2$ for ln BS of 24% & 22% respectively) (Table 2). The dispersion in our multi-level model cross-validation also compared well with this LUR study; for example RMSE values (in *ln* transformed units) of 0.300 and 0.419, respectively (or 18.8 and 6.8 μg $m^{-3}$, respectively, in BS concentration).

Another broadly comparable study modelled exposure to BS in the Netherlands between 1985-96 using LUR and GIS-based interpolation methods (Beelen et al., 2007). The overall models (involving regional, urban and local components) explained 59% of the variation in long-term average monitored concentrations, against which our multi-level model is again comparable. (Overall model performance statistics at independent 'hold out' test sites were not quoted by Beelen et al. (2007)). In NE England the model developed by Dadvand et al. (2011) explained 70% of spatio-temporal variation, although it appears that a large proportion of this variation may have arisen from temporal variables (independent 'hold out' model evaluations of long-term average spatial variations in BS were not quoted).

Epidemiological analyses with the exposure estimates produced in this work indicate that long-term exposure to air pollution is associated with increased risk of mortality (Yap et al., In press). This is consistent with other epidemiological research (Pope and Dockery, 2006). However, our epidemiological analyses were critically sensitive to the exposure assignment model used. IDWBS model estimates had much lower associations with mortality than those provided by the LAQP-based AMBS and MultiBS models, which is consistent with theoretical expectations about exposure misclassification (Sheppard et al., 2012).

In summary, improved modelling of exposure is essential to the quantification of the health impacts of long-term exposure to air pollutants and to inform public policy on future air quality standards. This study illustrates the salient challenges involved and presents pragmatic approaches to dealing with these. In particular it has highlighted the crucial importance of exposure model selection when determining associations between air pollution exposure and health outcomes.

8. Acknowledgements

We gratefully acknowledge funding from Department of Health (England) Policy Research Programme as part of the Initiative on Air Pollution (Research Grant 0020015). The views in this paper are those of the authors but not necessarily the Department of Health (England). The authors are also particularly grateful for the encouragement and support given by the late David Hole former Professor of Epidemiology and Biostatistics Head of West of Scotland Cancer Surveillance Unit in his enthusiastic role in planning and implementing this research.

References

Beelen, R., Hoek, G., Fischer, P., Brandt, P.A.v.d., Brunekreef, B., 2007. Estimated long-term outdoor air pollution concentrations in a cohort study. Atmospheric Environment 41, 1343-1358.

Beelen, R., Voogt, M., Duyzer, J., Zandveld, P., Hoek, G., 2010. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. Atmospheric Environment 44, 4614-4621.

Beverland, I.J., Cohen, G.R., Heal, M.R., Carder, M., Yap, C., Robertson, C., Hart, C.L., Agius, R.M., 2012. A comparison of short-term and long-term air pollution exposure associations with mortality in two cohorts in Scotland. http://dx.doi.org/10.1289/ehp.1104509. Environmental Health Perspectives.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. The Lancet 360, 1233-1242.

Dadvand, P., Rushton, S., Diggle, P.J., Goffe, L., Rankin, J., Pless-Mulloli, T., 2011. Using spatio-temporal modeling to predict long-term exposure to black smoke at fine spatial and temporal scale. Atmospheric Environment 45, 659-664.

DEFRA, 2005. National air quality data archive. Department for Environment Food & Rural Affairs (DEFRA) http://www.airquality.co.uk/ [accessed 1 October 2005].

DETR, 1999. Department of Environment Transport & Regions (DETR) Instruction Manual: UK Smoke and Sulphur Dioxide Network, AEAT-1806. AEA Technology, Harwell.

Dockery, D.W., Pope, C.A., III, Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., Jr., Speizer, F.E., 1993. An association between air pollution and mortality in six U.S. cities. New England Journal of Medicine 329, 1753-1759.

Durban, M., Harezlak, J., Wand, M.P., Carroll, R.J., 2005. Simple fitting of subject-specific curves for longitudinal data. Statistics in Medicine 24, 1153-1167.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V., Pearce, M.S., Parker, L., Charlton, M., Pless-Mulloli, T., 2008. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. Environmetrics 19, 549-566.

Gryparis, A., Coull, B.A., Schwartz, J., Suh, H.H., 2007. Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. Journal of the Royal Statistical Society Series C-Applied Statistics 56, 183-209.

Gulliver, J., Briggs, D., 2011. STEMS-Air: A simple GIS-based air pollution dispersion model for city-wide exposure assessment. Science of the Total Environment 409, 2419-2429.

Gulliver, J., Morris, C., Lee, K., Vienneau, D., Briggs, D., Hansell, A., 2011. Land Use Regression Modeling To Estimate Historic (1962-1991) Concentrations of Black Smoke and Sulfur Dioxide for Great Britain. Environmental Science & Technology 45, 3526-3532.

Hart, C.L., MacKinnon, P.L., Watt, G.C., Upton, M.N., McConnachie, A., Hole, D.J., Smith, G.D., Gillis, C.R., Hawthorne, V.M., 2005. The midspan studies. International Journal of Epidemiology 34, 28-34.

Heal, M.R., Quincey, P., 2012. The relationship between black carbon concentration and black smoke: A more general approach. Atmospheric Environment 54, 538-544.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561-7578.

Janssen, N.A.H., Hoek, G., Simic-Lawson, M., Fischer, P., van Bree, L., ten Brink, H., Keuken, M., Atkinson, R.W., Anderson, H.R., Brunekreef, B., Cassee, F.R., 2011. Black Carbon as an Additional Indicator of the Adverse Health Effects of Airborne Particles Compared with PM(10) and PM(2.5). Environmental Health Perspectives 119, 1691-1699.

Jerrett, M., Burnett, R.T., Ma, R.J., Pope, C.A., Krewski, D., Newbold, K.B., Thurston, G., Shi, Y.L., Finkelstein, N., Calle, E.E., Thun, M.J., 2005. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology 16, 727-736.

Krewski, D., Jerrett, M., Burnett, R.T., Ma, R., Hughes, E., Shi, Y., Turner, M.C., Pope, C.A., Thurston, G., Calle, E.E., Thun, M.J., 2009. Extended Follow-Up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality. HEI Research Report 140. Health Effects Institute, Boston, MA. http://pubs.healtheffects.org/view.php?id=315 Accessed 2012.

Maynard, D., Coull, B.A., Gryparis, A., Schwartz, J., 2007. Mortality risk associated with short-term exposure to traffic particles and sulfates. Environmental Health Perspectives 115, 751-755.

NAEI, 2005. National Atmospheric Emissions Inventory (NAEI): Annual average daily traffic flows at count points on major roads (Scotland). http://www.naei.org.uk/

Ordnance-Survey, 1993. PANORAMA DTM data set. http://edina.ac.uk/digimap/ Accessed: 2006.

Ordnance-Survey, 2003. Strategi data set (for urban edge). http://edina.ac.uk/digimap/ Accessed: 2006.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama-Journal of the American Medical Association 287, 1132-1141.

Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: Lines that connect. Journal of the Air & Waste Management Association 56, 709-742.

Pope, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E., Heath, C.W., 1995. Particulate Air-Pollution As A Predictor of Mortality in A Prospective-Study of Us Adults. American Journal of Respiratory and Critical Care Medicine 151, 669-674.

R-Development-Core-Team, 2006. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org: ISBN 3-900051-07-0.

Shaddick, G., Wakefield, J., 2002. Modelling daily multivariate pollutant data at multiple sites. Journal of the Royal Statistical Society Series C-Applied Statistics 51, 351-372.

Sheppard, L., Burnett, R., Szpiro, A., Kim, S.-Y., Jerrett, M., Pope, C., Brunekreef, B., 2012. Confounding and exposure measurement error in air pollution epidemiology. Air Quality, Atmosphere & Health 5, 203-216.

SURPOP, 2006. Modelled household count in 200 m grid squares from the 1981 census. Source: The 1991 Census, Crown Copyright. ESRC/JISC purchase. http://www.census.ac.uk/cdu/surpop/

Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 95-114.

Yanosky, J.D., Paciorek, C.J., Schwartz, J., Laden, F., Puett, R., Suh, H.H., 2008. Spatio-temporal modeling of chronic PM10 exposure for the nurses' health study. Atmospheric Environment 42, 4047-4062.

Yap, C., Beverland, I.J., Heal, M.R., Cohen, G.R., Robertson, C., Henderson, D.E.J., Ferguson, N.S., Hart, C.L., Morris, G., Agius, R.M., In press. Association between long-term exposure to air pollution and specific causes of mortality in Scotland. . Occupational and Environmental Medicine DOI:10.1136/oemed-2011-100600.

Table 1. Summary of Local Air Quality Predictor (LAQP) variables at monitoring sites.

| Variable | Mean | Median | Interquartile range | Min | Max |
|---|---|---|---|---|---|
| Easting | - | - | - | 2142 | 3645 |
| Northing | - | - | - | 6211 | 7410 |
| Altitude (m) | 59.3 | 44 | 17-89 | 1 | 248 |
| Household density (250 m buffer) | 250.2 | 197.6 | 84.9-348.0 | 0 | 1041.9 |
| Distance to nearest major road (km) | 0.467 | 0.221 | 0.101-0.548 | 0.002 | 4.26 |
| Distance to edge of urban boundaries (km) | 0.46 | 0.31 | 0.064-0.53 | –2.56 | 3.52 |

Table 2. Cross-validation statistics for inverse distance weighted (IDWBS), additive (AMBS) and multi-level (MultiBS) exposure models for 10 random selections of 19 test sites and 162 training sites according to model evaluation procedure outlined in section 2.4.

| Exposure model | $R^2$ (%) | RMSE | FB | SD FB | Bias | %.Bias | SD % Bias |
|---|---|---|---|---|---|---|---|
| (A) BS data (μg $m^{-3}$ units): | | | | | | | |
| IDWBS | 22 | 7.9 | 0.02 | 0.39 | –0.7 | 17 | 79 |
| AMBS | 24 | 7.9 | 0.01 | 0.39 | –1.0 | 16 | 83 |
| MultiBS | 46 | 6.8 | 0.08 | 0.30 | 1.2 | 14 | 37 |
| Gulliver et al. (2011) † | 40 | 18.8 | –0.09 | - | - | - | - |
| (B) *ln* transformed BS data: | | | | | | | |
| IDWBS | 19 | 0.413 | 0.02 | 0.16 | 0.025 | 4 | 21 |
| AMBS | 18 | 0.413 | 0.01 | 0.16 | 0.011 | 3 | 21 |
| MultiBS | 60 | 0.300 | 0.03 | 0.11 | 0.081 | 4 | 12 |
| Gulliver et al. (2011) † | 59 | 0.419 | - | - | - | - | - |

Footnotes:

† Mean of statistics quoted in this paper for modelled years 1971 and 1981.

$R^2$: explained variance for 190 pairs of predicted and observed exposures (from 10 random selections of 19 test sites).

RMSE: root mean squared error.

Fractional Bias: $FB = \dfrac{(C_p - C_o)}{0.5 \times (C_p + C_o)}$

where $C_p$ and $C_o$ are predicted and observed concentrations.

SD FB: standard deviation of the fractional bias.

$$\%\ \text{Bias} = \frac{(C_p - C_o)}{C_o} \times 100$$

SD % Bias: standard deviation of the percentage bias

The RMSE, SD FB and SD %Bias were calculated within each of the ten replications (as variances), averaged over the 10 replications and then square rooted.

The mean FB and % Bias were calculated as means within each of the 10 replications and then the average of each of the within replications means.

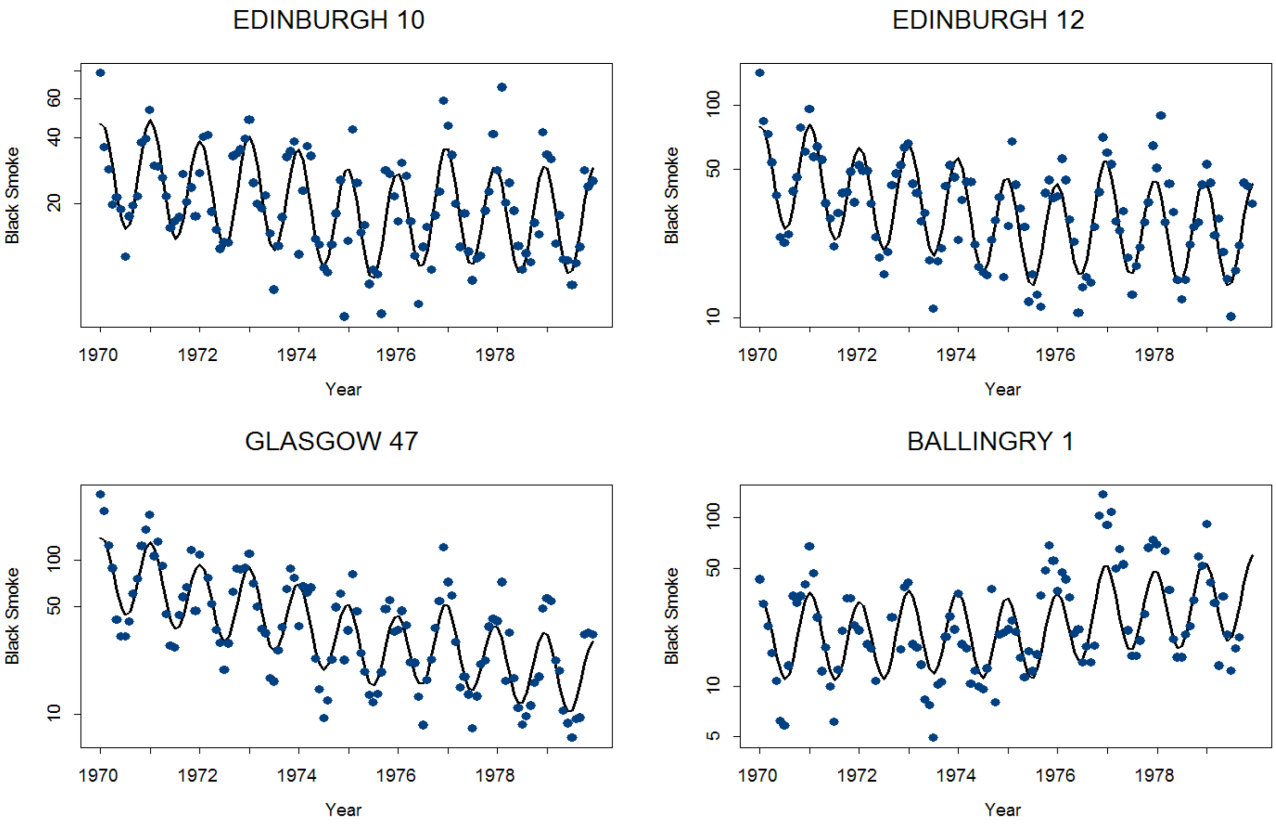
Figure captions (for figures on following 3 pages):

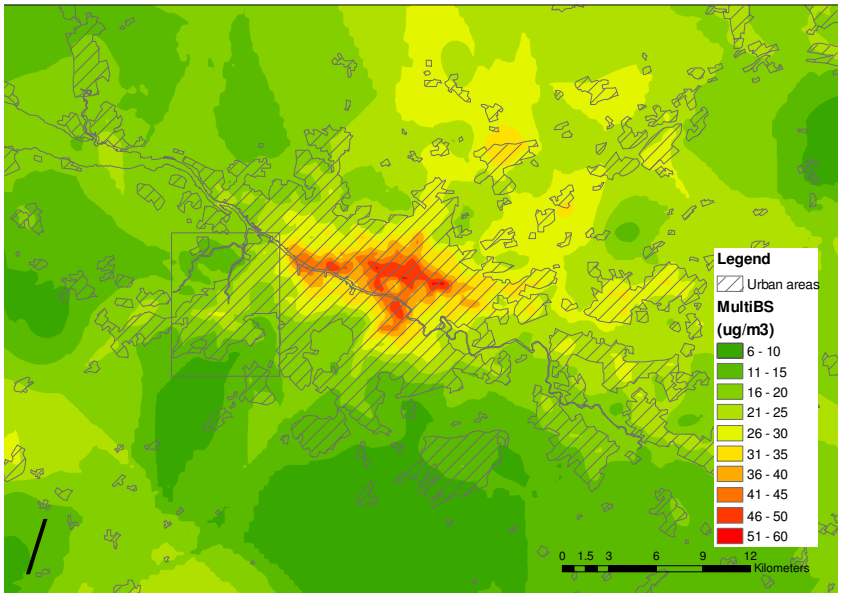
Figure 1. Multilevel model prediction of monthly black smoke ($\mu g\ m^{-3}$) at four measurement sites. Points are measurements.

Figure 2. Black smoke exposure estimates in Glasgow conurbation using 3 different exposure models. Exposures have been visualised by kriging exposure estimates at cohort individuals' postcode centroids. Glasgow conurbation is in the west part of central Scotland (see supplementary Fig S1 for more extensive land outline of Scotland) and is bisected by the River Clyde. Renfrew and Paisley cohort area (Figure 3) indicated by rectangle. Colour versions of this figure are available in the on-line version of this article.

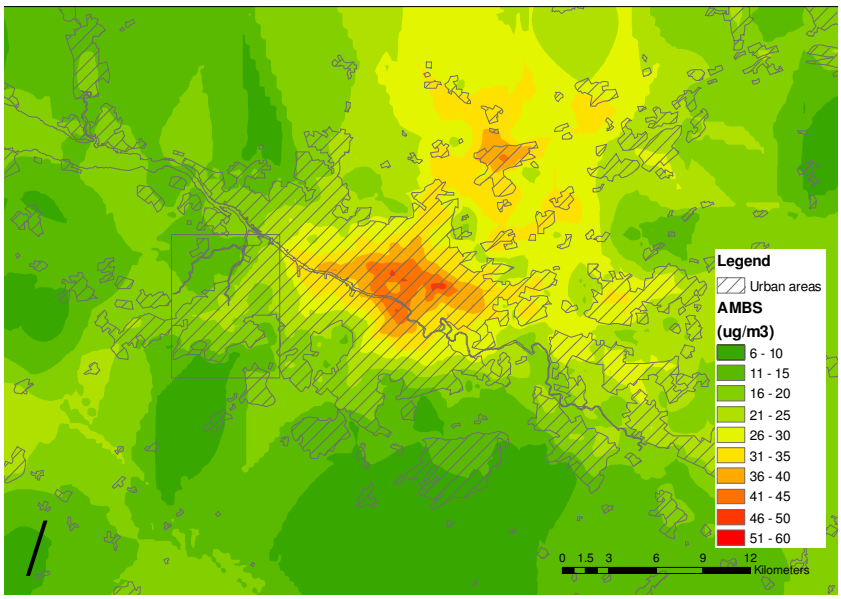
Figure 3. Black smoke exposure estimates in Renfrew and Paisley using 3 different exposure models. Exposures have been visualised by kriging exposure estimates at cohort individuals' postcode centroids (small diamonds). The geographical location of this area is shown by the rectangles in Figure 2. Colour versions of this figure are available in the on-line version of this article.

Figure 1

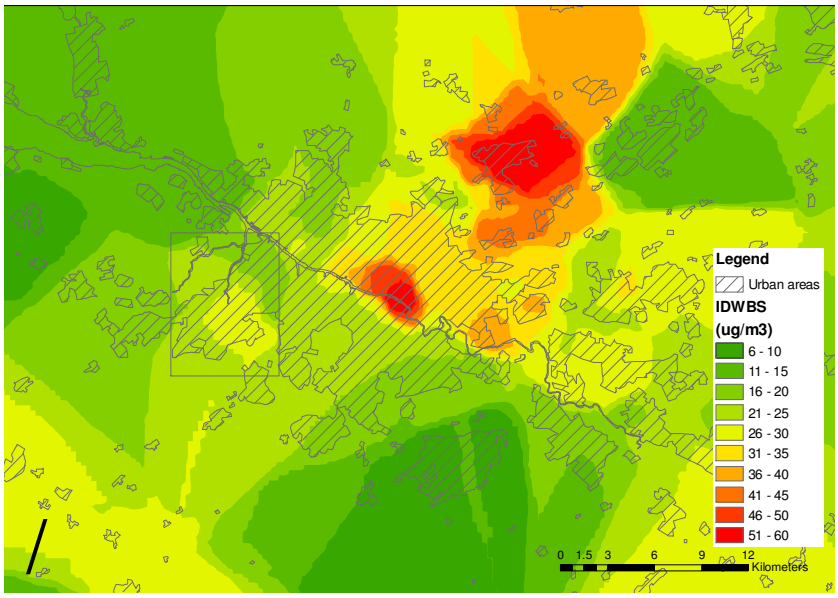

EDINBURGH 10
Black Smoke
20
40
60
1970
1972
1974
1976
1978
Year
EDINBURGH 12
Black Smoke
10
50
100
1970
1972
1974
1976
1978
Year
GLASGOW 47
Black Smoke
10
50
100
1970
1972
1974
1976
1978
Year
BALLINGRY 1
Black Smoke
5
10
50
100
1970
1972
1974
1976
1978
Year




(A) Multi-level model exposure estimates.

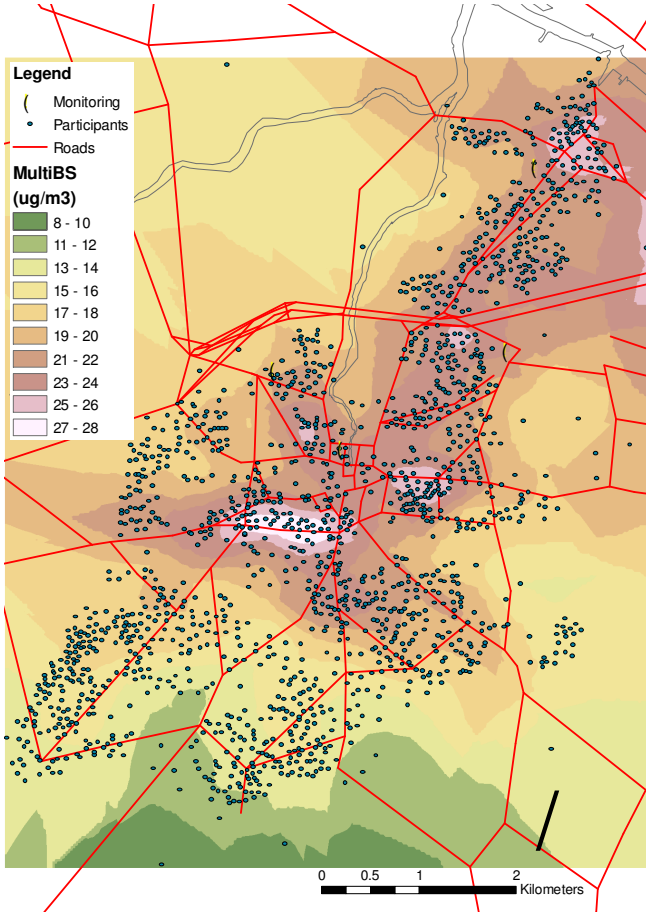



(B) Additive model exposure estimates.

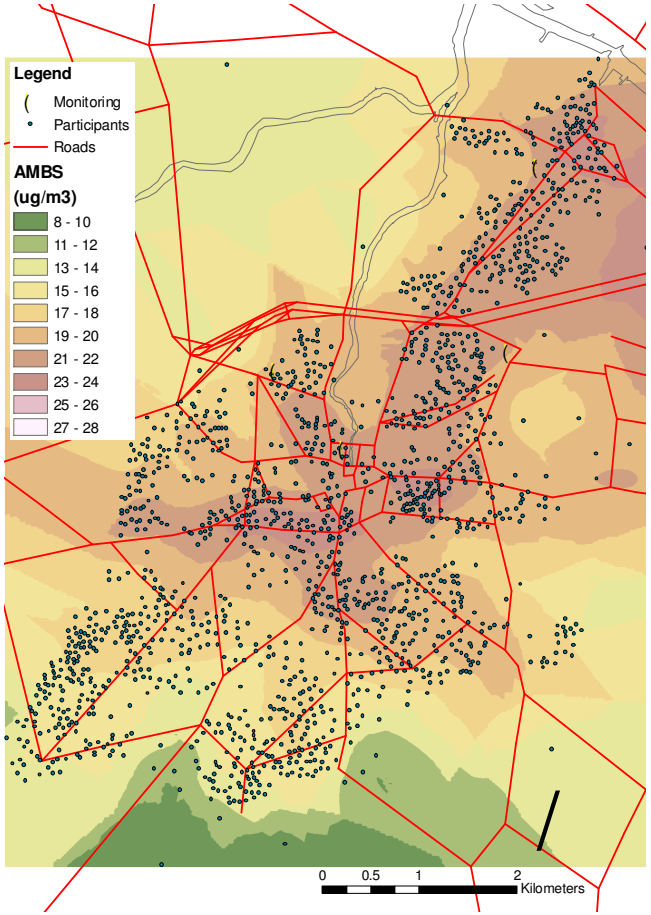



(C) Inverse distance weighted model exposure estimates.

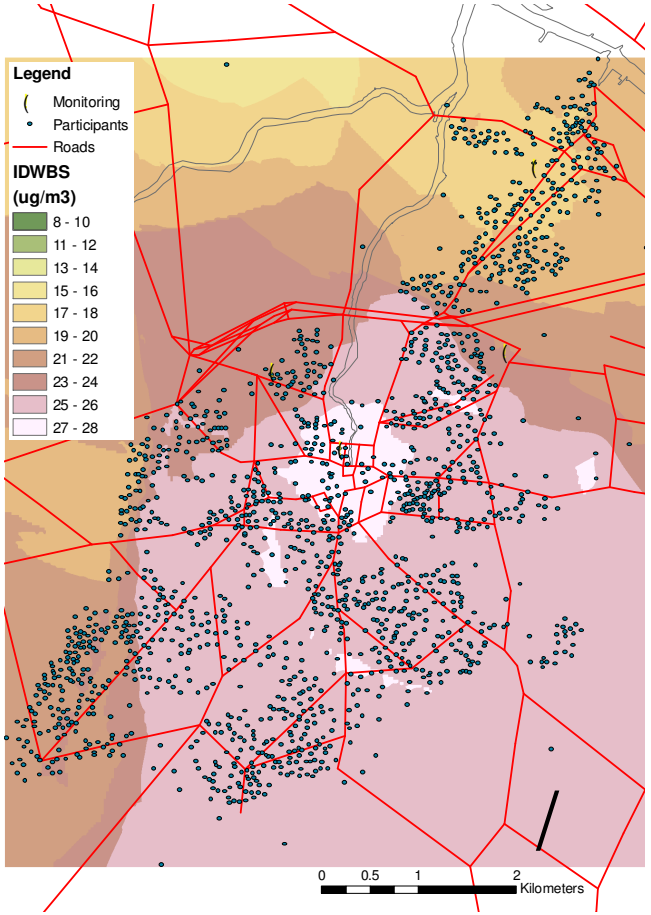
Figure 2




(A) Multi-level model prediction




(B) Additive model prediction




(C) Inverse distance weighted model prediction

Figure 3

# Supplementary information

## Comparison of Models for Estimation of Long-Term Exposure to Air Pollution in Cohort Studies

Beverland, I.J.*
*Department of Civil Engineering, University of Strathclyde, Glasgow, UK*

Robertson, C.
*Mathematics & Statistics, University of Strathclyde;*
*Health Protection Scotland, Glasgow, UK; and*
*International Prevention Research Institute, Lyon, France*

Yap, C.
*Department of Civil Engineering, University of Strathclyde, Glasgow, UK*
*(Current address: MRC Midland Hub for Trials Methodology Research,*
*University of Birmingham, UK)*

Heal, M.R.
*School of Chemistry, University of Edinburgh, Edinburgh, UK*

Cohen, G.R.
*Edinburgh, UK*

Henderson, D.E.J.
*Department of Civil Engineering, University of Strathclyde, Glasgow, UK*
(Current address: SKM Enviros, Shrewsbury, UK)

Hart, C.L.
*Institute of Health and Wellbeing, Public Health, University of Glasgow, Glasgow, UK*

Agius, R.M.
*Centre for Occupational and Environmental Health, The University of Manchester, UK*

***Correspondence to:**
Dr Iain Beverland,
Senior Lecturer,
University of Strathclyde, Department of Civil Engineering,
John Anderson Building, 107 Rottenrow,
Glasgow G4 0NG
Tel: 0141-548-3202
Fax: 0141-553-2066
Email: Iain.Beverland@strath.ac.uk

S1 Map of monitoring sites and cohort individual postcode centroids

A map showing the geographical locations of monitoring sites and cohort individual postcode centroids is given in Figure S1.

S2. Exemplification of missing pollution data

Data from the 5 monitoring sites around the Renfrew-Paisley cohort illustrate the common problem of substantial amounts of missing data (Figure S2, upper panel). The Paisley 8 site was only in operation for two periods early in the 1970s while the Renfrew 3 site operated continuously from 1972 until early 1979 and Linwood 1 operated from 1971 to 1979 with a gap in 1978. Paisley 5 and 7 operated throughout the whole period of interest but not continuously and in the mid 1970s only operated during the winter.

S3. Sensitivity analyses for area-based log linear regression and inverse distance weighted exposure model (IDWBS)

Sensitivity analyses involved: (a) allowing the linear and change point effects of time to vary across sites; (b) replacing the factors for the day of week and month of year effects with sine and cosine curves; (c) using piecewise linear trends; and (d) using regression-based multiple imputation methods in SOLAS (2006) and SPSS (2006). For the 5 sites within the Renfrew-Paisley Cohort area (within region 21 of Figure S1) the simple model (Equation 1 main paper) produced broadly equivalent estimates to the more complex models (Table S1). At the four monitoring sites where there was a substantial amount of data the different imputation methods tended to have similar imputed values (Table S1). The main differences observed were for the Paisley 8 site, for which data were only available for a short time, and there the site-specific slopes model produced quite different geometric mean concentrations compared to the other methods. Although this led to a significant site by time interaction in the model this was a result of this one monitoring site that had little data. As the monitoring sites were all geographically close to each other (within 10 km) we expected the same trends in all sites.

Our statistical modelling was informed by environmental knowledge and we pursued our analyses with the simpler model with no interaction in the other regions. Consequently, Equation 1 was applied to all 162 sites which had ≥ 10% of data available during 1970-79. Median and geometric means of the non-missing daily BS and the geometric means over all days in the period using the imputations from Model 1 were compared for the 30 sites which had the lowest number of missing

observations (Table S2). In all but two of the sites the geometric mean with imputations was slightly higher than the geometric mean without imputations, suggesting a slight bias. However, 14 of these sites had most of their missing data during the first 3 years and 18 had > 66% of their missing days during the winter. Both of these factors make it likely that imputation would raise the average exposure. BS concentrations tended to be higher during winter months as there was greater use of household coal fires then. Smoke control areas were introduced during the 1970s in many urban areas in Scotland and this led to a gradual reduction in BS concentrations.

S4 Local Air Quality Predictors (LAQP)

Maps illustrating examples of LAQP are shown in Figure S3.

S5 Sensitivity analyses for additive exposure model

A nonparametric bivariate smooth trend over spatial locations (Easting and Northing) and parametric or nonparametric terms for LAQP were used. The most general model fitted was:

$$\ln(BS + 0.5) = s(E, N) + s(A) + s(HD) + s(MR) + s(UB) + UA.Ind + \varepsilon \quad \text{(S1)}$$

The smooth trends were represented using penalised regression splines (Wood, 2000). Multidimensional smoothes used penalised thin plate splines (Wood, 2003). Parameters were estimated using Generalised Cross Validation (GCV), through selection of 'best fit' models from all possible values of the smoothing parameters (Wood, 2000; 2003; 2006). Linear relationships were observed between *ln* BS and altitude and household density. Nonlinear relationships were apparent for distance to the major roads and distance to the edge of the urban areas. Besides visualising the relationship between each predictor (individually and simultaneously) and BS by fitting a smooth curve, Box Tidwell additivity and variance stabilisation transformations were also used to select the most appropriate transformation for the spatial predictors (Box and Tidwell, 1962).

Five spatial regression models were examined (Table S3). These ranged from a fully flexible additive model to a linear regression model with no spatial smoothing. The most parsimonious model (Model 4) was a semi-parametric model with bivariate smooth trend for *s(E,N)* and parametric terms for the GIS-derived spatial predictors.

S6 Sensitivity analyses using different multi-level models

The following models were fitted to the data:

- Model ML1-I NSC: Equation 4 (main paper) with the four spatial covariates, $A_i$, $\sqrt{HD_i}$, $\sqrt{MR_i}$, and $UB_i$, removed and with $f(t_{ij})$ and $g_i(t_{ij})$ modelled using Equations 5 and 6, respectively
- Model ML1-I: Equation 4 with $f(t_{ij})$ and $g_i(t_{ij})$ modelled using Equations 5 and 6, respectively
- Model ML2-IT: Equation 4 with $f(t_{ij})$ and $g_i(t_{ij})$ modelled using Equations 5 and 7, respectively
- Model ML3-IT(NP): Equation 4 with $f(t_{ij})$ and $g_i(t_{ij})$ modelled using Equations 5 and 8, respectively
- Model ML3-IT(NP-Serial): ML3-IT(NP) with an additional term to model the serial correlation between data for successive months

In all of these models seasonal effects were included using a sine and cosine term to represent the months rather than 11 dummy variables. Parameter estimates for the models were obtained using REML (Restricted Maximum Likelihood). Likelihood ratio tests used maximum likelihood estimation. Adding the spatial predictors improved the fit of the model significantly (model ML1-I NSC *c.f.* ML1-I (Table S4)). There were also benefits in making the deviations from population trend more flexible (model ML1-I *c.f.* ML1-IT and ML1-IT *c.f.* ML3-IT(NP) (Table S4)). There was evidence that temporal correlation should be taken into account. Model ML3-IT(NP-Serial) had the highest log likelihood and the lowest AIC and BIC. In this model an exponential temporal correlation term was included. Hence the more flexible model ML3-IT(NP) Serial (using penalised splines for both the average trend and site-specific differences) was preferred (Table S5).

The parameter estimates for model ML3-IT(NP-Serial) showed that BS decreased with increasing altitude of a monitoring site and with increasing distance from a major road, and that BS increased with increasing household density and with proximity to an urban centre (Table S5). The seasonal variation in mean *ln* BS was greater than the variation associated with the LAQP (e.g. the increase of mean *ln* BS from the minimum (0) to the maximum (1042) of households within 250 m buffer zones was $0.015 \times \sqrt{1042} = 0.48$ units, which was about half as much as the increase from summer to winter). There was little difference in the estimated effects of the LAQP when using sines and cosines to represent the seasonal trend as compared to monthly indicators and the former model was used as fewer parameters were required. The random effect parameters indicated residual variation within sites of the same order of magnitude as variation in the intercepts across sites ($\sigma$ and $\sigma_0$). The correlation between observations in adjacent months was exp(-1/1.0076) = 0.37.

S7 Visualisation of estimated BS spatial variations for entire Collaborative cohort area:

Maps of predicted BS concentrations derived from the IDWBS, AMBS and MultiBS exposure models produced by spatial smoothing of the BS estimates from each model at the address postcodes of the 21,621 cohort individuals are shown in Figure S4. These maps were produced using a nonparametric bivariate smooth trend using penalised thin plate splines (Wood, 2003) fitted using the *mgcv* package in R (R-Development-Core-Team, 2006) and were consistent with those from an alternative use of kriging within ARCGIS in the Figures 2 and 3 of the main paper. Both examples of spatial smoothing illustrated were undertaken only for visualisation purposes; individual cohort participants have individual exposure estimates at their location.

References:


Box, G., Tidwell, P., 1962. Transformation of the independent variables. Technometrics 4, 531-550.

R-Development-Core-Team, 2006. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org: ISBN 3-900051-07-0.

SOLAS, 2006. SOLAS 3.0 for Missing Data Analysis. http://www.statsol.ie/html/solas/solas_home.html.

Splus, 2006. Splus 7.0 professional statistical program. http://www.insightful.com/products/splus/default.asp.

Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society Series B-Statistical Methodology 62, 413-428.

Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 95-114.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC. ISBN 9781584884743.

Table S1. Sensitivity of imputations in the Renfrew-Paisley Cohort area. Estimated decadal geometric mean black smoke (μg $m^{-3}$) with missing daily data replaced by predicted values from different imputation models.

| Site | Missing[a] | LLR1[b] | LLR1s[c] | LLR2[d] | LLR2s[e] | Solas[f] | SPSS[g] |
|---|---|---|---|---|---|---|---|
| Paisley 5 | 919 | 27.6 | 27.7 | 27.6 | 27.7 | 28.5 | 29.2 |
| Paisley 7 | 1058 | 22.6 | 22.5 | 22.6 | 22.5 | 23.2 | 23.9 |
| Paisley 8 | 3220 | 12.1 | 7.3 | 12.2 | 6.9 | 11.5 | 23.0 |
| Renfrew 3 | 1098 | 17.1 | 17.7 | 17.1 | 17.7 | 17.1 | 16.3 |
| Linwood 1 | 902 | 16.8 | 16.4 | 16.7 | 16.4 | 17.0 | 16.5 |

[a] The number of days with missing data. There are 3652 days in the study period.
[b] Equation 1 (main paper) LLR – Log Linear Regression.
[c] Equation 1 with the piecewise linear trend varying across sites.
[d] Equation 1 with the dummy variables for day of the week and month replaced by cosine and sine terms of period 7 days and 12 month, respectively.
[e] Equation 1 with the dummy variables for day of the week and month replaced by cosine and sine terms of period 7 days and 12 month, respectively, and an interaction with site on the piecewise linear trend.
[f] Multiple imputation using regression imputation in SOLAS, mean of 5 sets of imputations.
[g] Multiple imputation using regression imputation in SPSS, mean of 5 sets of imputations.

Table S2. Observed and 'observed plus imputed' black smoke concentrations (μg $m^{-3}$) at the 30 monitoring sites with the lowest number of missing observations over the 10 year period 1970-79.

| Sites | % missing days | median[a] | geometric mean[a] | geometric mean[b] |
|---|---|---|---|---|
| EDINBURGH.10 | 1.2 | 18 | 18.94 | 19.01 |
| EDINBURGH.12 | 1.4 | 29 | 30.44 | 30.57 |
| GLASGOW.47 | 1.5 | 31 | 34.09 | 34.48 |
| EDINBURGH.14 | 1.6 | 31 | 31.52 | 31.64 |
| EDINBURGH.17 | 1.8 | 21 | 21.29 | 21.32 |
| EDINBURGH.20 | 1.8 | 41 | 40.06 | 40.15 |
| EDINBURGH.16 | 2.4 | 19 | 18.97 | 19.06 |
| GLASGOW.60 | 2.5 | 27 | 27.98 | 28.30 |
| ESKDALEMUIR.1 | 2.6 | 3 | 3.41 | 3.43 |
| GLASGOW.44 | 2.6 | 23 | 24.21 | 24.41 |
| GLASGOW.61 | 2.7 | 15 | 16.23 | 16.33 |
| GLASGOW.52 | 3.0 | 45 | 47.30 | 48.19 |
| GLASGOW.62 | 3.6 | 21 | 23.55 | 23.80 |
| EDINBURGH.18 | 3.6 | 24 | 24.17 | 24.20 |
| LANARKSHIRE.11 | 3.9 | 35 | 37.74 | 38.24 |
| GRANGEMOUTH.2 | 4.0 | 14 | 15.45 | 15.67 |
| GLASGOW.66 | 4.2 | 20 | 20.15 | 20.52 |
| GLASGOW.51 | 4.8 | 22 | 24.28 | 24.44 |
| EDINBURGH.15 | 5.2 | 29 | 30.45 | 30.45 |
| STIRLING.COUNTY.8 | 7.1 | 16 | 16.87 | 17.34 |
| GLASGOW.42 | 7.9 | 24 | 25.78 | 26.83 |
| BALLINGRY.1 | 8.5 | 22 | 22.65 | 22.75 |
| CLYDEBANK.5 | 8.6 | 14 | 14.01 | 13.88 |
| LANARKSHIRE.15 | 9.7 | 16 | 17.00 | 16.83 |
| PORT.GLASGOW.1 | 10.4 | 14 | 14.66 | 15.06 |
| GLASGOW.67 | 10.5 | 16 | 16.77 | 17.54 |
| CLYDEBANK.6 | 10.6 | 22 | 22.38 | 22.46 |
| GLASGOW.68 | 11.2 | 28 | 30.35 | 32.98 |
| PORT.GLASGOW.4 | 11.3 | 10 | 10.24 | 10.54 |
| COCKENZIE.5 | 12.0 | 10 | 10.87 | 11.00 |

[a] Observed data only without any imputation over the 10 year period from 1970 - 79.
[b] Combined observed data and missing data imputed from Equation 1 (main paper).

Table S3. Local Air Quality Predictor regression models.

| Model | Formulae for linear predictor | Model type [a] | $R^2$ adj | Df | AIC | BIC |
|---|---|---|---|---|---|---|
| 1 | $\beta_1 A + \beta_2 \sqrt{HD} + \beta_3 \sqrt{MR} + \beta_4 UB + UA.Ind$ | *A* | 0.35 | 6 | 207.5 | 225.9 |
| 2 | $s(E) + s(N)$ | *B ii* | 0.44 | 15.3 | 194.7 | 244.9 |
| 3 | $s(E,N)$ | *B ii* | 0.59 | 25.0 | 154.0 | 233.7 |
| 4 | $s(E,N) + \beta_1 A + \beta_2 \sqrt{HD} + \beta_3 \sqrt{MR} + \beta_4 UB + UA.Ind$ | *B i* | 0.75 | 30.1 | 77.2 | 172.8 |
| 5 | $s(E,N) + s(A) + S(HD) + S(MR) + S(UB) + UA.Ind$ | *B ii* | 0.79 | 42.6 | 61.2 | 194.8 |

[a] *A* Linear regression models (parametric),
[a] *B* Generalised additive models, *i* semi-parametric, *ii* nonparametric.
*s(x)* A nonparametric smooth function of the predictor *x*, that is estimated by smoothing splines
*E* Easting
*N* Northing
*A* Altitude
*HD* Household density within 250 m buffer
*MR* Distance to nearest major road
*UB* Shortest distance to edge of urban area
*UA.Ind* Indicator for within a small or large urban area, or outside of an urban area
$R^2$ adj adjusted R-square
Df Total estimated degrees of freedom used
AIC Akaike information criterion
BIC Bayesian information criterion (tends to penalise models with more degrees of freedom than the AIC)

AIC, BIC and adjusted $R^2$ are model comparison criteria. They are commonly used to compare and select the most parsimonious model, which should have the lowest AIC and BIC but highest adjusted $R^2$.

Table S4. Multilevel model testing effects of the four local air quality predictors.

| Model | P | AIC | BIC | log_Lik | LR | *p* |
|---|---|---|---|---|---|---|
| ML1-I NSC | 7 | 11590 | 11641 | -5788 | | |
| ML1-I | 11 | 11513 | 11594 | -5746 | 84.2 | < 0.0001 |
| ML2-IT | 13 | 10593 | 10688 | -5284 | 924.2 | < 0.0001 |
| ML3-IT(NP) | 14 | 10543 | 10646 | -5258 | 51.8 | < 0.0001 |
| ML3-IT(NP-Serial) | 15 | 9211 | 9320 | -4591 | 1334.7 | < 0.0001 |

P: Number of parameters estimated
AIC: Akaike information criterion
BIC: Bayesian information criterion
log_Lik: log Likelihood
LR: log Likelihood Ratio Test Statistic

In model ML1-I NSC there are 4 fixed effects, the intercept, the linear trend and the sine and cosine seasonal effects. There are also the variances of the 3 random effects corresponding to the site specific intercept, $\sigma_V^2$, the change in slope at the knots, $\sigma_u^2$, and the within site error term, $\sigma^2$.

Table S5. Parameter estimates from the multilevel model ML3-IT(NP-Serial) (Section S6 supplementary data) with sine/cosine or indicator terms for season.

| | Sine/cosine curves | | | Monthly indicators | | |
|---|---|---|---|---|---|---|
| Fixed Effects | Estimate | SE | *p* | Estimate | SE | *p* |
| *Intercept* | 3.5411 | 0.1055 | <0.001 | 4.0141 | 0.1114 | <0.001 |
| *Time* | -0.0011 | 0.0039 | 0.774 | -0.0035 | 0.0027 | 0.197 |
| *Jan* | | | | 0 | - | |
| *Feb* | | | | 0.0906 | 0.0165 | <0.001 |
| *Mar* | | | | -0.2983 | 0.0164 | <0.001 |
| *Apr* | | | | -0.5181 | 0.0165 | <0.001 |
| *May* | | | | -0.7061 | 0.0164 | <0.001 |
| *Jun* | | | | -1.0032 | 0.0165 | <0.001 |
| *Jul* | | | | -1.1828 | 0.0165 | <0.001 |
| *Aug* | | | | -0.8834 | 0.0166 | <0.001 |
| *Sep* | | | | -0.6737 | 0.0165 | <0.001 |
| *Oct* | | | | -0.2355 | 0.0165 | <0.001 |
| *Nov* | | | | -0.0599 | 0.0164 | 0.003 |
| *Dec* | | | | 0.0199 | 0.0165 | 0.226 |
| *Season - Cosine* | 0.5585 | 0.0066 | <0.001 | | | |
| *Season - Sine* | 0.0729 | 0.0067 | <0.001 | | | |
| *A* | -0.0014 | 0.0006 | 0.017 | -0.0016 | 0.0007 | 0.015 |
| $\sqrt{HD}$ | 0.0150 | 0.0046 | 0.002 | 0.0145 | 0.0051 | 0.005 |
| $\sqrt{MR}$ | -0.3093 | 0.0835 | <0.001 | -0.3196 | 0.0924 | 0.001 |
| *UB* | 0.1838 | 0.0420 | <0.001 | 0.2176 | 0.0453 | <0.001 |
| *Random Effects* | Estimate | L95 | U95 | Estimate | L95 | U95 |
| $\sigma_V$ | 0.0009 | 0.0003 | 0.0029 | 0.0011 | 0.0005 | 0.0024 |
| $\sigma_0$ | 0.4932 | 0.4257 | 0.5714 | 0.4947 | 0.4278 | 0.5722 |
| $\sigma_1$ | 0.0039 | 0.0030 | 0.0051 | 0.0038 | 0.0029 | 0.0050 |
| $\rho_{01}$ | -0.575 | -0.718 | -0.385 | -0.590 | -0.732 | -0.398 |
| $\sigma_u$ | 0.0231 | 0.0139 | 0.0385 | 0.0214 | 0.0129 | 0.0355 |
| *Correlation range* | 1.0076 | 0.9402 | 1.0799 | 1.0053 | 0.9716 | 1.0402 |
| $\sigma_\varepsilon$ | 0.3790 | 0.3723 | 0.3858 | 0.3592 | 0.3540 | 0.3645 |

*A*: Altitude
*HD*: Household Density within 250 m buffer
*MR*: Distance to nearest major road
*UB*: Shortest distance to edge of urban area
L95: Lower 95% confidence limit
U95: Upper 95% confidence limit

$\sigma_V$ *:* Standard deviation of the site specific random effect at each spline knot

$\sigma_0$ *:* Intercept standard deviation, element of Σ

$\sigma_1$ *:* Trend standard deviation, element of Σ

$\rho_{01}$ *:* Correlation between intercept and trend random effects, element of §

$\sigma_u$ *:* Standard deviation of the random effect of the spline trend

$\sigma_\varepsilon$ *:* Within-site standard deviation

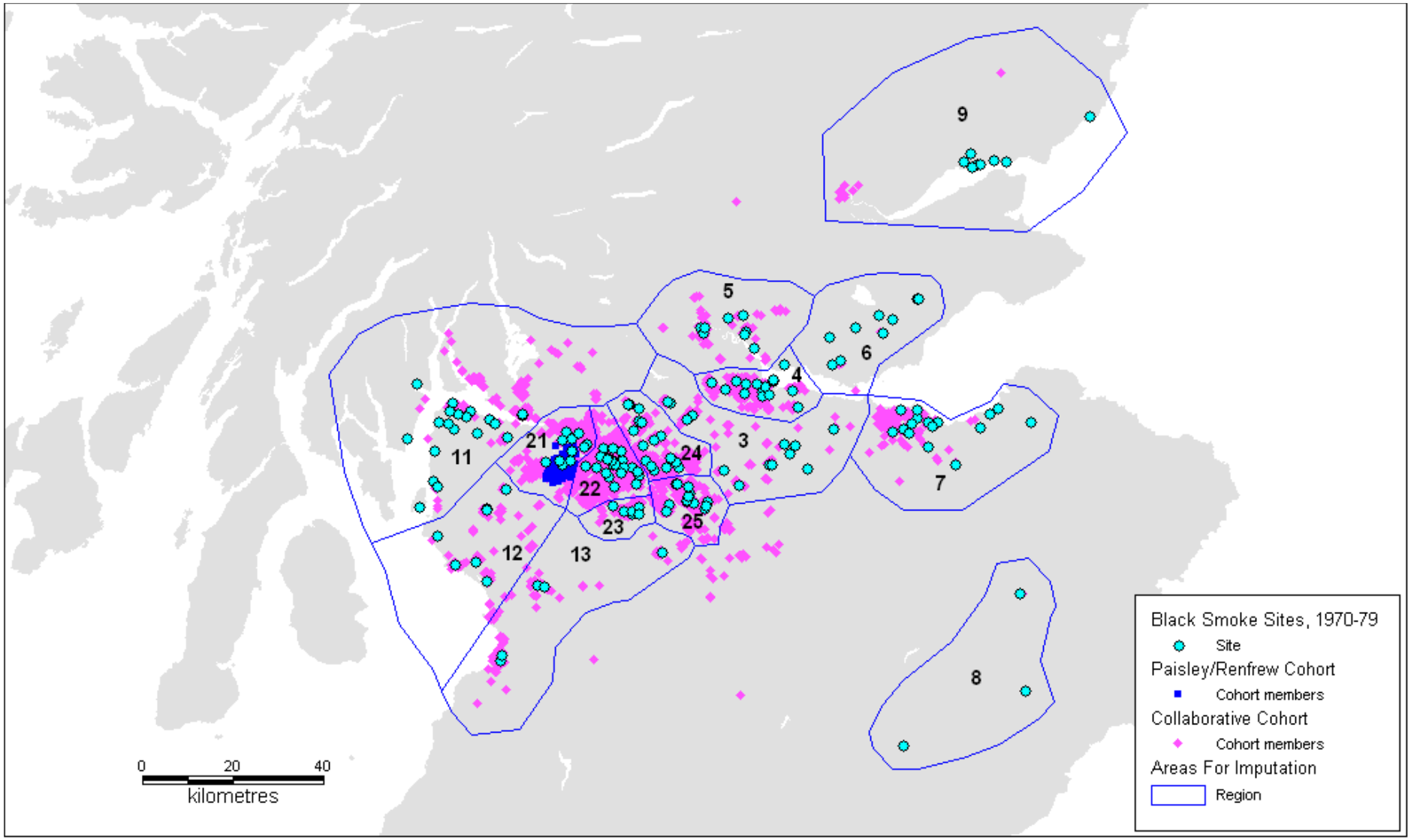



Figure S1. Locations of cohort individuals and black smoke monitoring sites illustrating regions used for area-based log linear regression model. Note that the 15 regions are labelled: 11,12,13 (areas west of Greater Glasgow conurbation); 21, 22, 23, 24, 25 (Greater Glasgow Conurbation); 3 (West Lothian); 4 (Falkirk & Grangemouth); 5 (Stirling); 6, (West Fife); 7 (Edinburgh conurbation); 8 (Borders); 9 (Tayside).

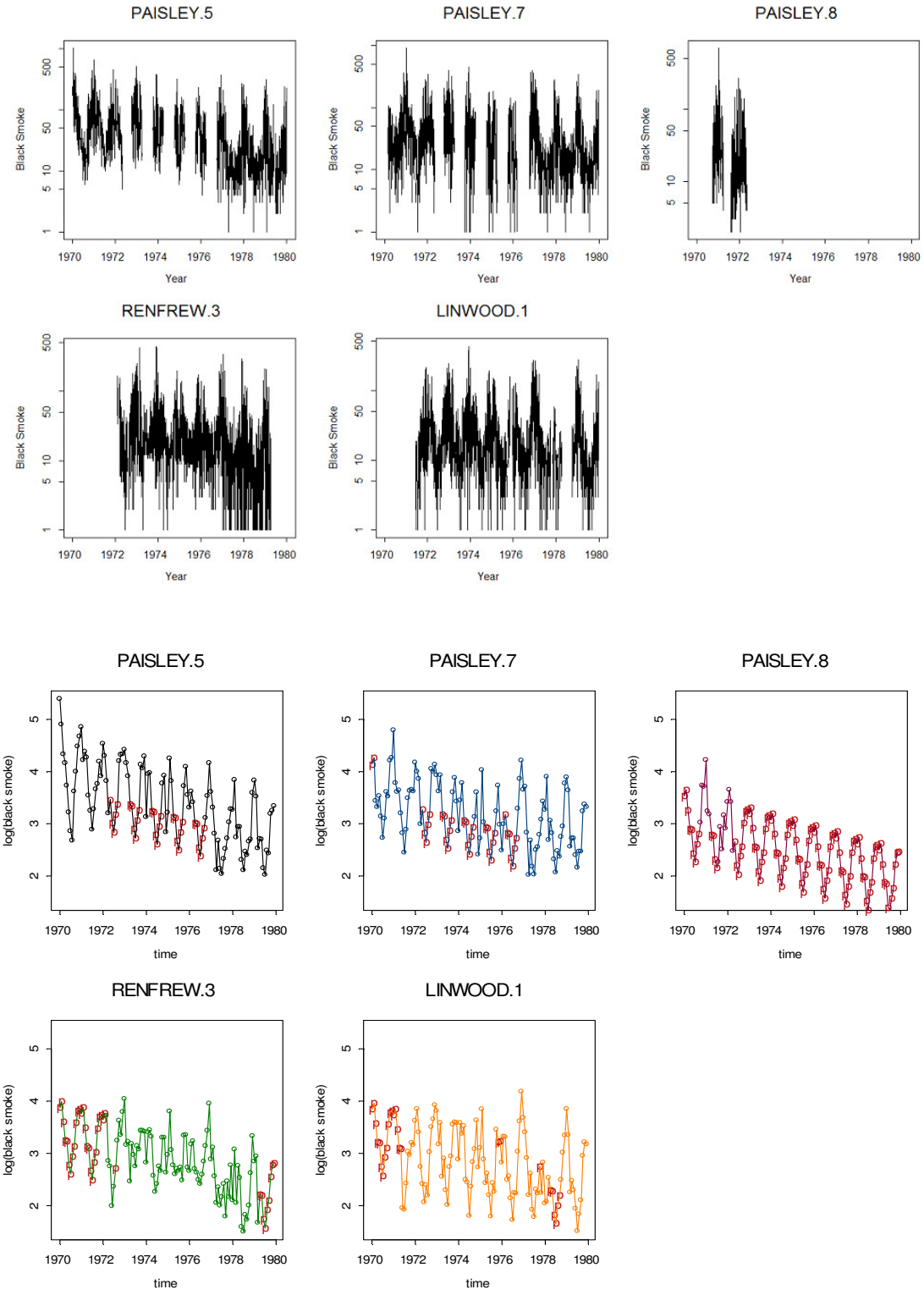
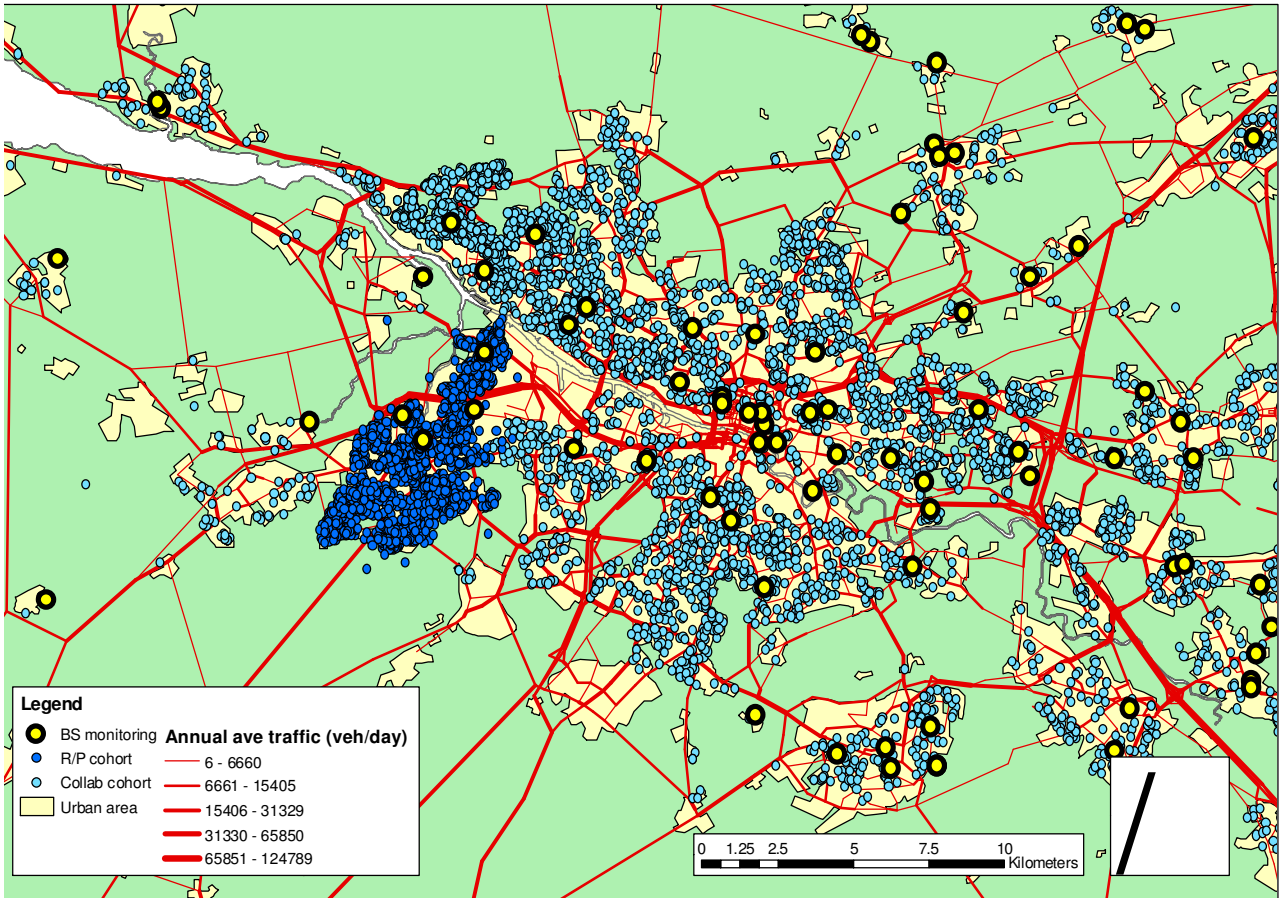
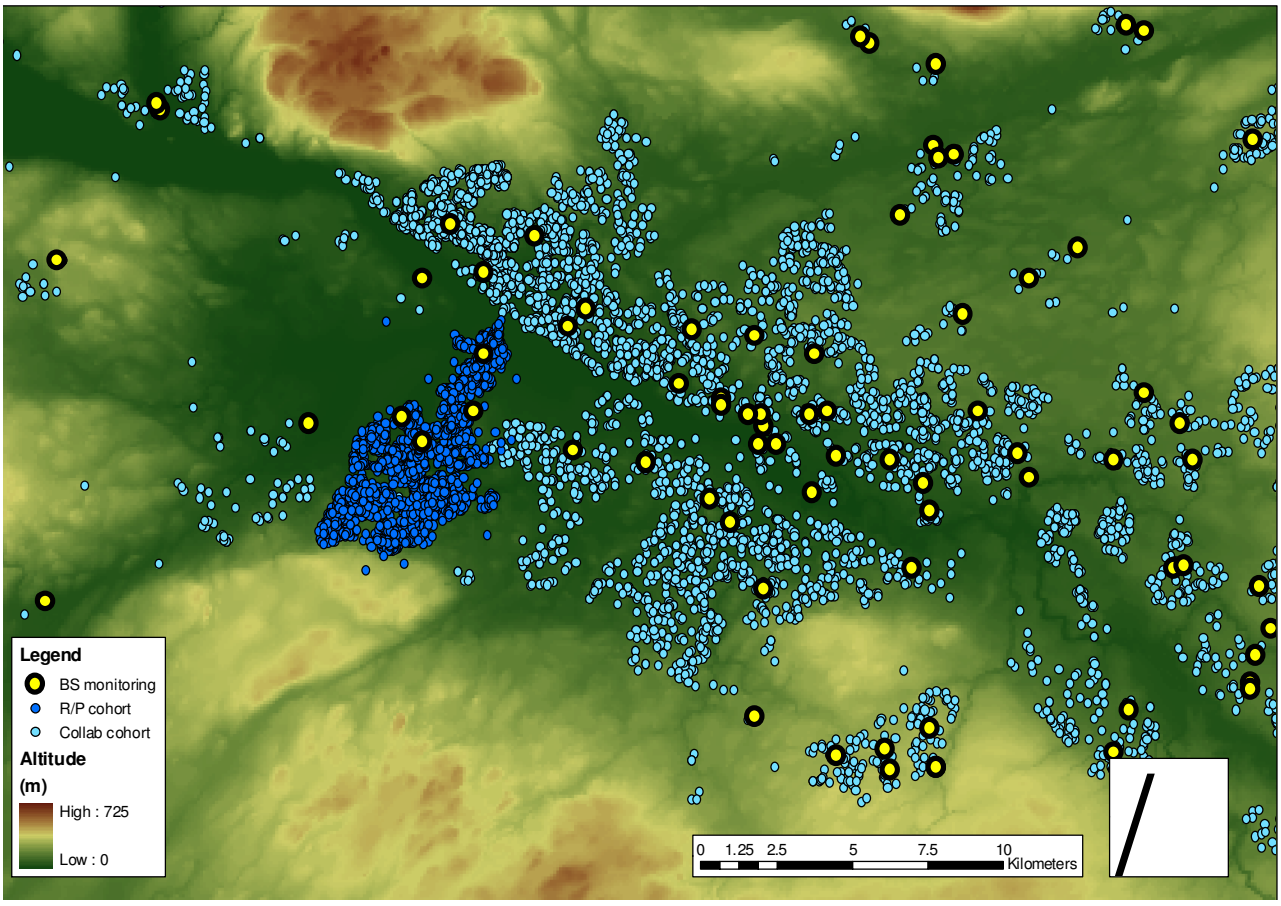



Figure S2. Upper panel: observed monthly black smoke concentrations (µg $m^{-3}$) at five monitoring sites in the Renfrew-Paisley Cohort area. Lower panel: observed and area-based log linear model imputed monthly *ln* geometric mean concentrations (p denotes imputed values and o denotes observations).

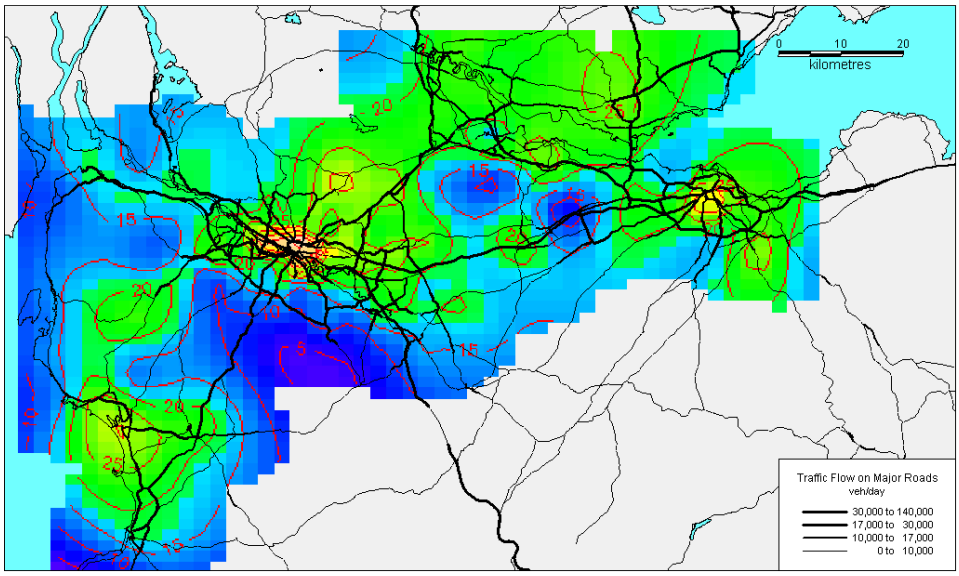



(B) Major roads and defined urban areas for cohort residence locations. NB the illustrated traffic flows were not used in the exposure models (see section 2.3.3)

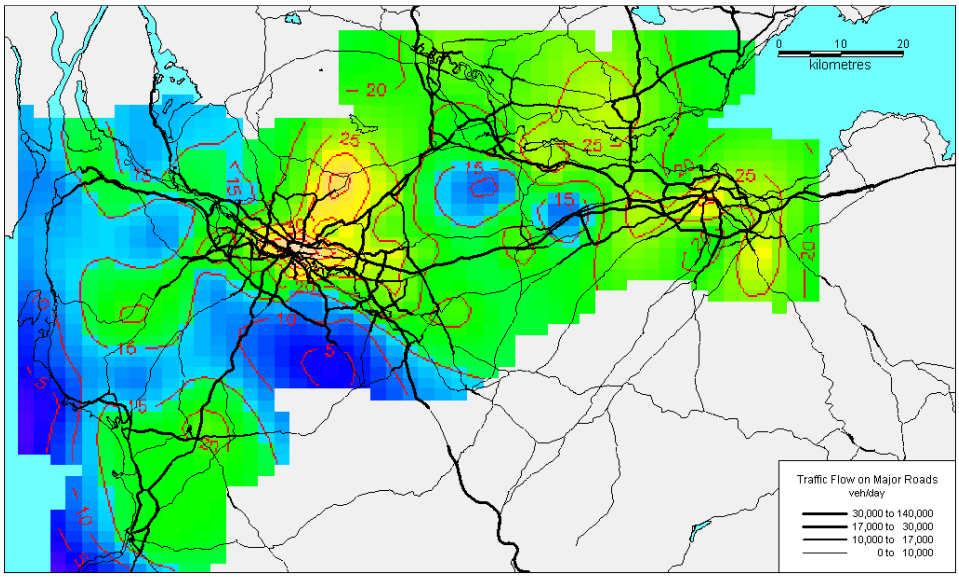



(A) Altitude of terrain at and surrounding cohort residence locations

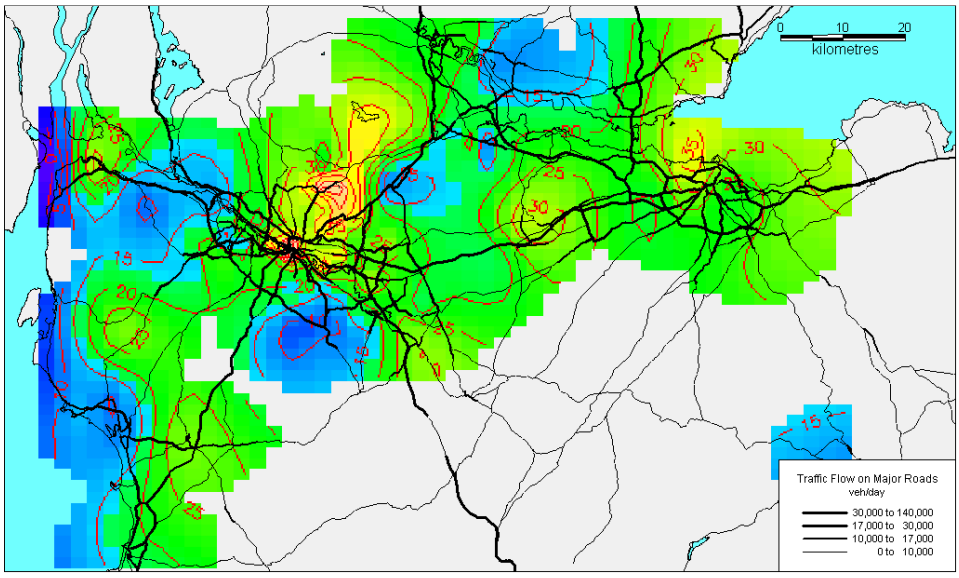
Figure S3. Examples of geographical variation in local air quality predictors (LAQP) in the vicinity of individual cohort residence locations.




(A) Predicted exposure contours from multilevel model.




(B) Predicted exposure contours from spatial additive model.




(C) Predicted exposure contours from inverse distance weighted model.

Figure S4. Predicted 1970-79 geometric mean black smoke exposure contours (μg $m^{-3}$) across area of all cohort individual residence locations. Blue-Green-Yellow-Brown colours and contour lines indicate increasing concentrations. Predicted exposures at cohort individuals' postcode centroids have been smoothed with a nonparametric bivariate trend over spatial locations (Easting & Northing) using penalised thin plate splines fitted using the *mgcv* package in R (R-Development-Core-Team, 2006).